Задачи NLP
NLU

Виды задач NLP

NLP

- Optical Character Recognition
- Automatic Speech Recognition
- Tokenization
- Lemmatization
- Morphoparsing
- Dependency and Constituency Parsing
- …

NLU

- Machine Translation
- Text Categorization
- Information Extraction
- Natural Language Inference
- Argument Mining
- Summarization
- Simplification
- Dialogue Systems
- …

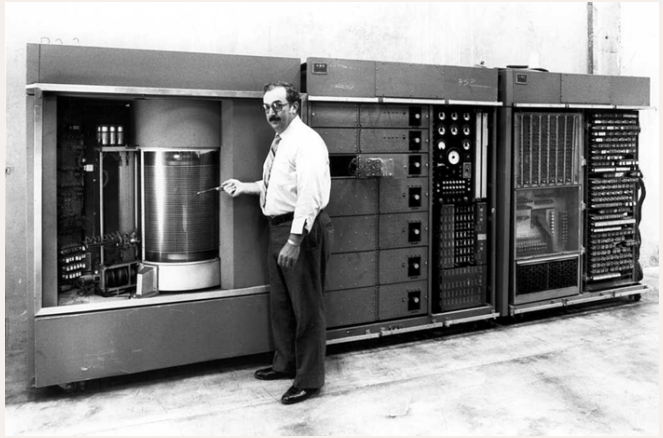
NLG

- (Machine Translation)
- (Summarization)
- (Dialogue Systems)
- Style Transfer
- Text Generation
- Text to Speech
- Text to Image
- Text to Video
- И наоборот

Машинный перевод

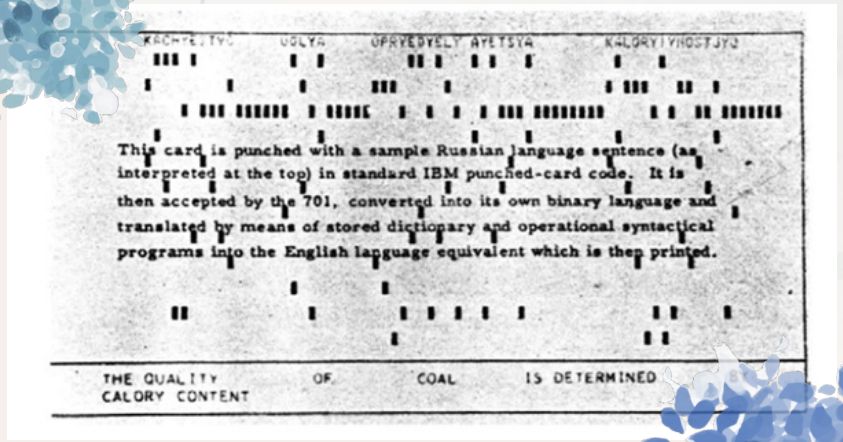
1947 год, письмо Уоррена Уивера Норберту Винеру:

I have wondered if it were unthinkable to design
a computer which would translate. Even if it would translate only scientific material
(where the semantic difficulties are very notably less), and even if it did produce
an inelegant (but intelligible) result, it would seem to me worth while.



Джорджтаунский эксперимент

- 1954 год
- Полностью переведены 60 предложений с русского на английский
- Предложения надо было записывать на перфокарты
- Система содержала всего шесть грамматических правил и
  250 лексических единиц в своем словаре
- Цель – ~~добыть у правительства побольше де-~~впечатлить публику



Перевод на правилах

- 60-е годы – пик интереса; но многие стали сомневаться в возможности
- В СССР – А.А. Ляпунов
- Первые системы переводили пословно
- Потом стали пытаться парсить синтаксис: такие системы называются Т-системы (transfer)
- В России ради машинного перевода разрабатывалась модель «Смысл-Текст» И.А. Мельчука

Перевод на правилах: сегодня

ЭТАП

- ЭлектроТехнический Автоматический Перевод
- Разрабатывается в ИППИ им. Харкевича с 1982 года
- Главный автор – И. Богуславский
- Последняя версия – 4
- Сегодня используется для теоретических исследований и автоматической разметки

Compreno

- Разрабатывался в компании ABBYY с начала нулевых
- Сегодня заморожен
- Но разработанный для него формат разметки используется для теоретических исследований
- Лежит в базе формата CoBaLD

Статистический машинный перевод

- Стали появляться большие параллельные корпуса
- Есть три вида МП на корпусах:

1. **МП на примерах (Example based)** – сегодня только в качестве вспомогательного; *Reverso context*
2. **Статистический МП** – классические алгоритмы пытаются найти закономерности
3. **МП на нейронных сетях** – то же, но нейронные сети

EMNLP и статистический МП

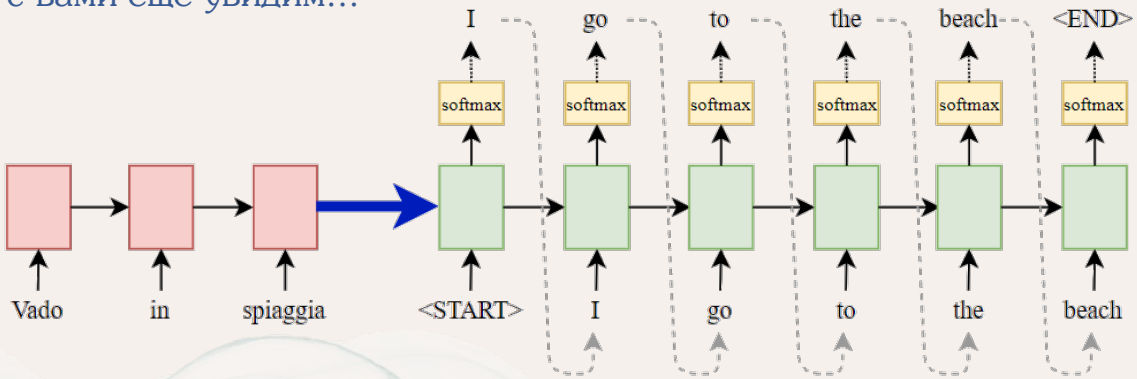
- 1996 год – первая конференция Empirical Methods in NLP
- **Одна из самых престижных NLP-конференций сегодня**
- 1999 – система статистического перевода Egypt
- 2006 – Google запустил свой Google Translate на статистическом МП
- В основе статистического МП – N-грамные языковые модели
- Считаем, что *любое предложение* языка X может быть с некоторой вероятностью переводом предложения языка Y. Наша задача – максимизировать вероятности для «хороших» вариантов

МП на нейронных сетях

- В основе МП на нейронной сети – **языковая модель** (сегодня – трансформер)
- 2014 год: Илья Сатскивер с группой коллег предложил метод МП на нейронных сетях
- Google перешел на нейронные сети в 2016 году

МП на нейронных сетях

- Классический подход: **энкодер** извлекает признаки (~семантическую информацию) из предложения языка X, а **декодер** на основе данных, полученных от энкодера генерирует новое предложение на языке Y
- Эту картинку мы с вами еще увидим…



Text Categorization

- Много разных видов задач
- Например, Sentiment Analysis
- Или классификация по темам или жанрам
- Или кластеризация – с темами и жанрами обычно работает хорошо
- Англоязычная модель на huggingface
- Мультиязычная модель на huggingface

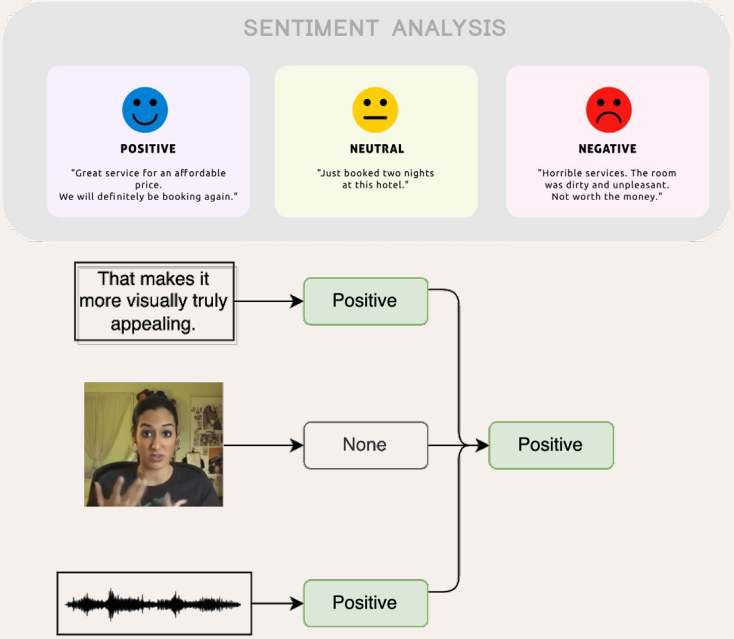
Sentiment Analysis

Анализ тональности текста – тоже важная бизнес-задача

Разновидности и SOTA-решения по ним включают:

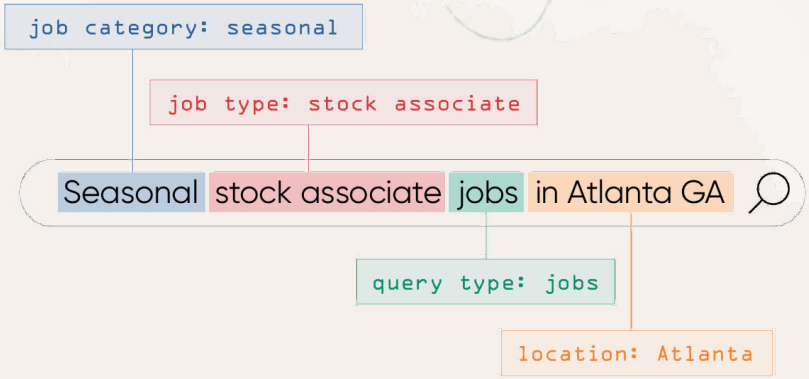
- emotion detection (Acheampong et al. 2021)
- multimodal sentiment analysis (Abdu et al. 2021; Chandrasekaran et al. 2021)
- cross-domain sentiment analysis (Al-Moslmi et al. 2017)
- multilingual sentiment analysis (Agьero-Torales et al. 2021)
- aspect-based sentiment analysis (Do et al. 2019; Nazir et al. 2020)
- subjectivity detection (Chaturvedi et al. 2018)

- *Для базового случая точность моделей приближается к 100% (95-98), поэтому интересны более сложные варианты*



Information Extraction

- Создание баз знаний (knowledge base)
- Named Entity Recognition and Classification
- Entity Linking
- Relationship Extraction



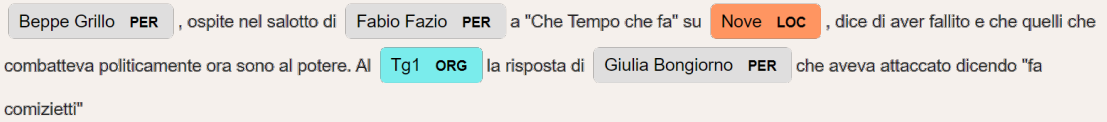
MUC

MUC-6

- Message Understanding Conferences: серия конференций, посвященных задаче извлечения информации
- На шестой конференции MUC (1995) предложили термин Named Entity Recognition and Classification
- А также классификацию именованных сущностей
- Помимо прочего, именно на MUC-6 сформулировали задачи coreference resolution, word sense disambiguation и predicate-argument structure parsing

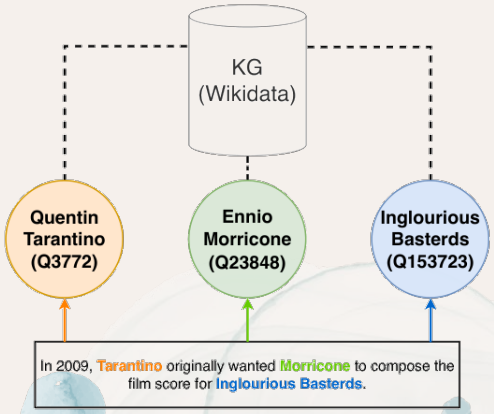
NER(C)

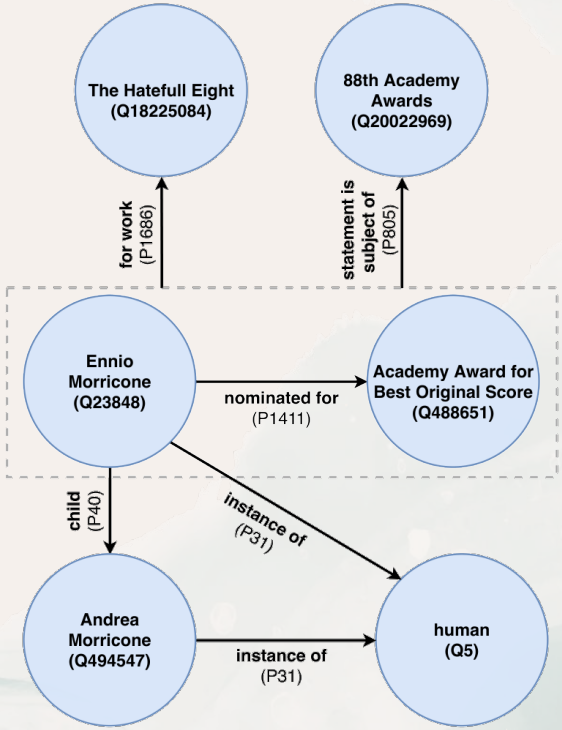
- Впервые попытались решить в 1991 году (Lisa Rau)
- Сперва решали эвристиками и ручными правилами
- **Газетир** – список именованных сущностей
- Сегодня решают с помощью нейронных сетей
- Классификации:
  - MUC: person, location, organization (enamex)
  - ACE: person, facility (location+organization), GPE (geopolitical entity)
  - CONLL: enamex + miscellaneous, иногда product
  - timex: time, date
  - numex: money, percent



Entity Linking

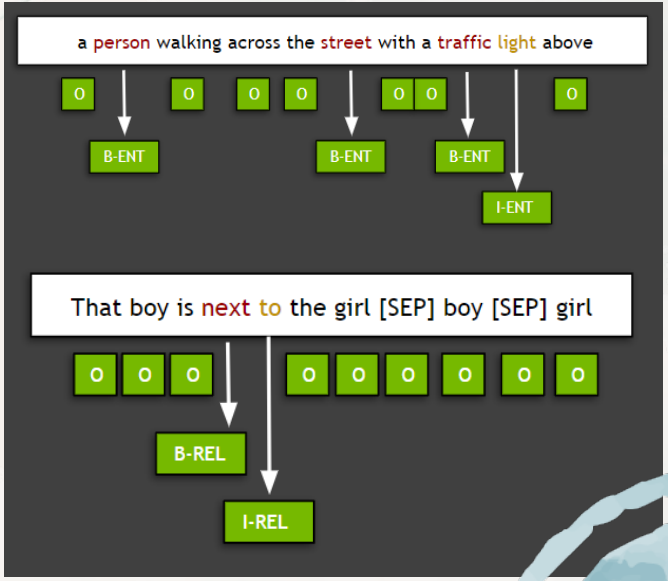
- Уже есть именованные сущности – хотим установить отношения между ними и собрать их в базу данных (граф данных, knowledge graph)
- Часто используют для обучения базы данных:
  - DBPedia
  - Freebase
  - Wikipedia
  - Wikidata





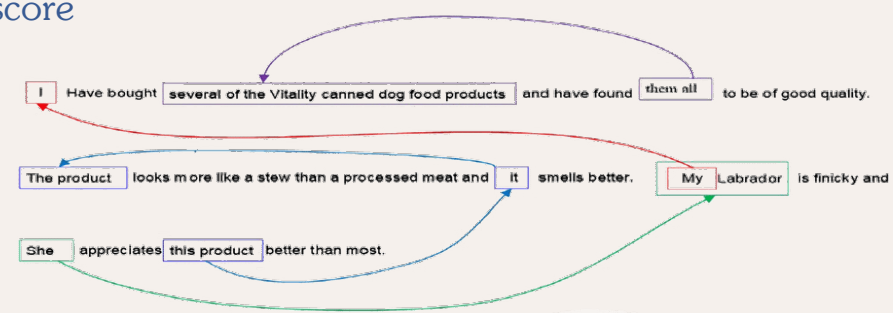
Relationship Extraction

- Устанавливаем связи между сущностями (необязательно именованными)
- Очень важная задача для бизнеса и обработки документов
- Иногда объединяют с задачей NER (EL тоже объединяют)



Coreference Resolution

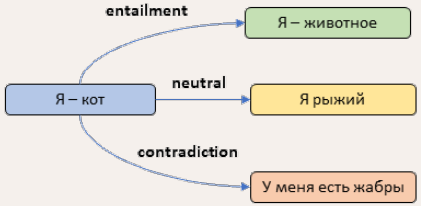
- Разрешение анафоры – важная задача для автоматического понимания текста
- Winograd schema challenge: «ноутбук не помещается в чемодан, потому что он маленький»
- Много сложностей: н-р, расщепленные антецеденты («*Борис Джонсон* женился на *госпоже Саймондс*. **Пара** поехала в свадебное путешествие»)
- Что вообще считать за сущность?
- Очень сложная задача: на соревновании RuCoCo в 2023 году для русского языка победитель достиг 75% F1-score



Natural Language Inference

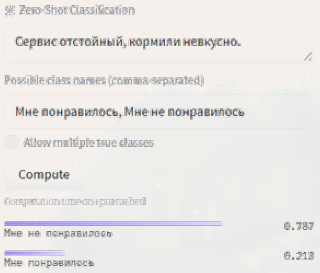
- Задача автоматического определения логической связи между текстами
- для двух утверждений A и B надо выяснить, следует ли B из A
- потыкать

Natural language inference – автоматическое определение логической связи между текстами



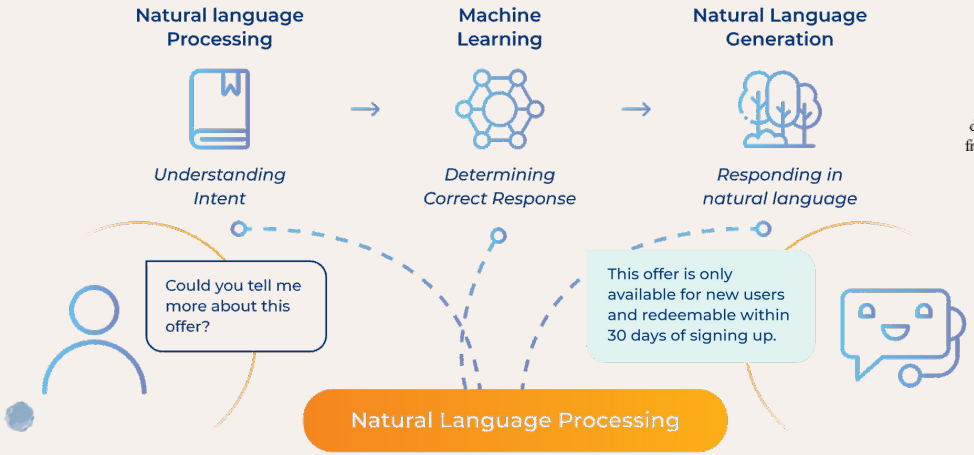
Теперь это могут и русскоязычные нейросети!

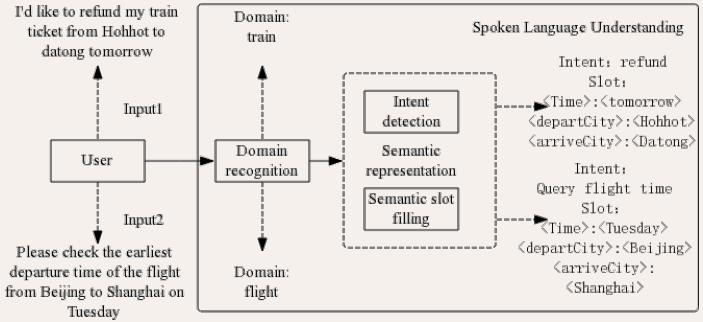
Пример применения NLI для zero-shot классификации тональности



Intent Detection and Slot Filling

- Детекция намерения: понять, чего хочет человек
- Заполнение слотов: извлечь семантические составляющие и занести в табличку





потыкать

Argument Mining

Text	Masks		Quarantine		Vaccines	
	Stance	Premise	Stance	Premise	Stance	Premise
И какой смысл в вакцине если антитела только 3 месяца? *(And what's the point of a vaccine if the antibodies work only for 3 months?)*	—	—	—	—	against	against
Должна быть вакцина которую, будут прививать с детства!!! *(There must be a vaccine that will be vaccinated from childhood!!!)*	—	—	—	—	for	no argument

Topic: Death Penalty

It does not deter crime and it is extremely expensive to administer .

Topic: Gun Control

Yes , guns can be used for protection but laws are meant to protect us , too.

- Задачи – извлечь из неструктуризованного текста гипотезу, позицию автора и его аргументы
- Аргумент обычно делится на позицию автора (claim), предпосылки (premises) и отношение между ними (support/attack links)
- Существует несколько датасетов разной сложности, качество сильно зависит от датасета (на некоторых – только 60-65%, на других до 80%)
- В 2022 году проводилось соревнование RuArg по AM для русского языка. Датасет был посвящен теме COVID. F1-мера победителей: по обнаружению позиции – 69,7%, по классификации предпосылок – 74%.
- потыкать

Word Sense Disambiguation

- Даже самым продвинутым архитектурам нейронных сетей (BERT, GPT) не так хорошо дается дизамбигуация: различение омонимов и разных значений многозначных слов
- Современные алгоритмы могут использовать вручную размеченные датабанки, н-р, WordNet или VerbNet
- Или семантическую разметку типа CoBaLD: по существу, парсер для этой разметки решает задачу all-word WSD
- Удивительно хорошо: обученный на очень маленькой модели ru-BERT tiny CoBaLD-парсер дает точность 90% по разметке семантических классов (=смыслов слов)
- Соревнование на SemEval-2023: Visual WSD (необходимо сопоставить одну из картинок со словом согласно его значению)
- Результаты победителей приближались к 90% на некоторых датасетах

Summarization

- Модель должна «прочитать» длинный текст, извлечь из него основные идеи и сделать краткое содержание
- Используется также генерация текста
- потыкать (англ)
- потыкать (мульти)